DOCUMENT RESUME

ED 308 006 PS 018 080

AUTHOR Walker, Richard N.
TITLE The Gesell Screening Examination: Psychometric Properties.
INSTITUTION Gesell Inst. of Human Development, New Haven, CT.
PUB DATE 89
NOTE 40p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Chronological Age; Early Childhood Education; Followup Studies; *Interrater Reliability; *Kindergarten Children; Longitudinal Studies; *Preschool Children; *Screening Tests; Sex Differences; Tables (Data); *Test Reliability; Test Validity
IDENTIFIERS *Developmental Screening; *Gesell Preschool Examination (Revised); Sociometric Status; Test Retest Reliability

ABSTRACT

In an assessment of the adequacy of the Gesell screening examination as a test instrument, a Gesell Screening Evaluation was given to 400 children semi-annually from their 4th to 6th year. The sample, which was stratified by parent occupation, included 40 girls and 40 boys at 5 age levels. The test battery corresponded with the Gesell Preschool and Kindergarten Screening Evaluations. Four experienced judges gave global developmental ratings and a grade placement recommendation for each child. For three judges, mean developmental age scores fell below chronological age by an interval of 2-7 months. For composite scores, mean inter-judge correlations by age groups ranged from .74 to .85 for single-judge ratings, and .85 to .92 for two-judge averages. Mean sex differences were small but significant, all favoring girls. The exam showed no differences by socioeconomic level. At 8 years, 182 children were retested, on multi-trait, multi-method measures. Screening ratings correlated most highly with later adaptive ability, and lower with number, fine motor, language and reading abilities, and general school competence. Lowest correlations were with gross motor and personal-social development. Suggestions for using Gesell screening evaluations to make placement recommendations are offered. (RH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED308006

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
X This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

The Gesell Screening Examination:

Psychometric Properties

Richard N. Walker

Gesell Institute of Human Development

New Haven, Connecticut

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Richard N. Walker
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Running head: GESELL SCREENING EXAMINATION

PS 018330

I thank Gail Bjorkland, Jackie Haines, Jan Keirns, Beth Lamb, Susan Landisio, and Gwen Webb for serving as raters.

Requests for reprints should be sent to Richard N. Walker, Gesell Institute of Human Development, 310 Prospect Street, New Haven, Connecticut 06511. (203) 777-3481.

BEST COPY AVAILABLE

Abstract

A Gesell Screening Evaluation was given to 400 4- to 6-year-olds--40 girls and 40 boys at five age levels, stratified by parent occupation. Four experienced judges gave global developmental ratings and a grade placement recommendation for each child. For three main judges, mean developmental age scores fell below chronological age by from 2 to 7 months; distributions suggested a different model than the bell-shaped one. For composite scores, mean inter-judge correlations by age groups ranged from .74 to .85 for single-judge ratings, from .85 to .92 for two-judge averages. Mean sex differences were small but significant, all favoring girls. The exam showed no differences by socio-economic level.

At age 8, 182 of the children were retested, on multi-trait, multi-method measures. Highest correlations of screening ratings were with later Adaptive ability; lower ones appeared with Number, Fine motor, Language, and Reading abilities and general School Competence. Lowest, but still significant, were correlations with Gross motor and Personal-Social development.

The Gesell Screening Examination: Psychometric Properties

From the time of his appointment in 1911 as director of the Yale Psycho-Clinic, Arnold Gesell was interested in supplementing naturalistic observations of infants and children in home-like settings with observations in more standard situations, facing simple tasks. His comparison of different children's reactions to the same tasks eventually became a method of assessing children.

The form of the assessment, its contents, and methods of recording and scoring have varied, depending in part on the colleagues working with Gesell. In 1925 and 1938 he gave detailed statistical presentations of distributions of children's responses on a number of items, in the latter book comparing three scoring methods for summarizing maturity status (Gesell, 1925; Gesell & Thompson, 1938). The First Five Years of Life (Gesell et al 1940) and Developmental Diagnosis (Gesell & Amatruda, 1941) also gave percentage distributions of subjects by age for test items, and presented a summarizing Developmental Schedule as a graphic display of the child's accomplishments in four areas of development (Motor, Adaptive, Language, and Personal-Social) by key ages.

However statistical his approach became, however systematic his scoring schema, Gesell maintained the position that his procedure was an examination, an observational procedure involving judgment by an experienced examiner, rather than a test, to be administered mechanically. Many psychometric questions important to researchers were never considered by name in his texts; the terms reliability and validity do not appear in

any of his indexes, for example. That is true also for the extension of Gesell's preschool exam for use with older children (Ilg & Ames, 1964), for a recent revision of the preschool exam (Ames, Gillespie, Haines, & Ilg, 1979), and for the shortened exam used in screening children for school entrance (Haines, Ames, & Gillespie, 1980; Ilg, 1982).

The abbreviated exam is widely used in assessing developmental maturity of children entering school, with a view to finding appropriate classroom placements for them. Its use is taught primarily in workshops. While many teachers and parents have found this testing and its recommendations to be helpful, administrators and psychologists have had unanswered questions about the documentation of developmental screening. Two main issues arise in evaluating approaches proposed by Gesell Institute staff. One concerns the results--short-term and long-term--of various attempts to match children's school settings to their developmental status. The other concerns the adequacy of the Gesell screening exam as a test instrument. This paper addresses the latter issue.

A problem with validation has been the question of the criterion: by what outcome measure could one judge the correctness of scores or placement recommendations from the screening tests? Such measures as later teacher ratings or reading scores or Gesell test ratings or grade retentions all offer relevant information but not definitive answers to questions about "hit rate." The present paper reports findings for a stratified sample of 4- to 6-year-old children on a test battery that substantially corresponds with the Gesell Preschool and Kindergarten Screening Evaluations. It describes distributions of scores by age, sex, and socio-economic status,

intercorrelation of scores, and reliability findings for the screening battery, and shows relationships with a revised version of the full Gesell Preschool Examination. It also offers information relevant to validity by a multi-trait, multi-method approach: a follow-up battery of tests and ratings of many of the subjects when they reached age 8.

Method

Subjects

The participating children were the 400 4- to 6-year-olds from a sample of children aged 2 1/2 to 6, tested in the 1970s. A first report on their testing appeared in The Child from One to Six (Ames, Gillespie, Haines, & Ilg, 1979). Their tests were used again in development of a Revised Gesell Preschool Examination (Walker, in press). Of these 400 children first seen as preschoolers, 182 were retested as 8-year-olds.

Eighty children--40 boys and 40 girls--were tested at each half-yearly age, within a month of the yearly or half-yearly birthday. At each age children were selected to fill an eight-level schedule of socio-economic status (SES) based on father's occupation. The sample was a New England one, most of the children living in Connecticut. For many children, race was not recorded; by estimate, between 5 and 7% of children at each age were nonwhite.

Table 1 shows the distribution of paternal occupations at every age, compared with 1970 U.S. Census data for adult employed males. Also shown are the paternal occupations for the subsample of 182 children who were retested at age 8. These "8-year-olds" actually ranged from 7.5 to 9.6

years (mean, M, = 8.6; standard deviation, SD, = 0.5 years). Results

---

Insert Table 1 about here

---

of the screening tests were not shared with parents or schools; children's eventual grade placements were based on their schools' customary policies.

The Revised Gesell Preschool Examination (RGPE) provides standard scores (M = 50, SD = 10) on five scales. When the children retested as 8-year-olds first took the exam, as 4- to 6-year olds, their mean score was 51.0 on the Motor scale, 50.9 on Adaptive, 50.8 on Verbal, 51.1 on Number, and 50.5 on Personal-Social. SDs ranged from 9.5 to 9.9. Both by fathers' occupations and by their own preschool scores, the retested children showed little difference from the children who were not retested.

The records described above allowed investigation of inter-judge agreement but not other forms of reliability. For a glimpse of test-retest reliability, an additional sample of 30 4-year-olds was tested (mean age of 51.7 months, SD of 3.3 months). These 17 girls and 13 boys, tested at several nursery school and day care centers, were retested an average of five weeks later (range of 20 to 81 days).

Measures

Screening exam. Three tests from the Adaptive section of the preschool exam form the nucleus of the screening examination: Copy Forms, Incomplete Man, and Cubes. Three brief verbal tests are often also given: Interview, Interests, and Naming Animals. They are not given much weight in the final evaluation and were not included in the present study.

Unlike most screening tests, which aim at simplicity in administering, scoring, and interpreting of items, the Gesell screening exam calls for assignment of global developmental ratings of each test, based on detailed familiarity with many age-specific characteristics of items (for example, how triangles are typically drawn at different ages). A composite developmental score is then given and an appropriate class placement is recommended.

Revised Gesell Preschool Examination. The RGPE gives age-equivalent and standard scores on Motor, Adaptive, Verbal, Number, and Personal-Social scales. Its norms are based on the sample described above, plus children similarly chosen at ages 2 1/2 to 3 1/2. Its reliability is shown in median alpha coefficients, for ages 4 to 6, of .90 for the Motor scale, .84 for Adaptive, .90 for Verbal, .94 for Number, and .76 for Personal-Social.

Though the screening evaluation tests and those in the RGPE Adaptive scale are mostly the same, their scoring differs in the two exams. The RGPE presents items by half yearly age levels to be scored as present or absent (+ or -). Its items are relatively simply defined--for example, number of missing parts added to the Incomplete Man. In the screening exam, in assigning a single developmental rating to the Incomplete Man, an experienced examiner probably considers about 200 items stored in her memory, each item keyed to a particular developmental age (for example, arm: style, length, angle, point of attachment, penetration, stroke, etc.).

Outcome measures. Individual measures obtained at age 8 included the Wide Range Achievement Tests of reading, spelling and arithmetic (Jastak, Bijou, & Jastak, 1965), school grade, teachers' ratings of 11 variables,

four tests from the Gesell Developmental Exam (Copy Forms, Incomplete Man, Naming Animals, and Monroe Visual III), an ad hoc test of gross and fine motor abilities, and a parental questionnaire of social maturity items, from the Gesell tests and the original Vineland Social Maturity Scale (Doll, 1965). All were scored as standard scores with age partialled out.

These 22 follow-up variables were grouped to give eight general variables, six intended to match Gesell's preschool domains--Fine and Gross Motor, Adaptive, Language, Number, and Personal-Social--plus Reading and a cluster rating general school Competence (attention span, ideal grade, actual grade for age, etc.). Each general variable was made up of scores on two to five measures, usually from different sources (test score, teacher rating, parent rating, school administrative action). Standard scores on the component measures were averaged in forming the summary scores, and the averages were then restandardized. The general scores were moderately intercorrelated among themselves. (See Walker, 1989.)

Procedure

From each child's full set of preschool exam records, the recording sheets for Copy Forms, Incomplete Man, and Cubes, and the child's drawings for Copy Forms and Incomplete Man were put together to comprise a screening exam. No scoring by the original examiner was noted on these records.

Five experienced developmental examiners--the main judges--first judged these test records. Judges A to E knew age and sex of each child and saw all three tests at the same time. They worked independently, and did not know SES or outcome findings for any children. Each judge rated all the records in three of the five age groups; each child's record was

rated by three judges. (Table 2 shows ages scored by each judge.) A sixth, particularly experienced judge later rated all the children's records. Judge F knew the mean scores that the main judges had collectively given to entire age groups; she did not know their ratings of individual children nor, of course, the children's outcome scores.

For each test and for the composite score, judges made developmental ratings on a scale graded by quarter years. In 15 records the Cubes test was poorly recorded and the composite score was based on the other two tests. Though in practice, recommendation of an appropriate classroom setting depends on grades available, and may also be achieved through mixed-age or nongraded settings, for this research judges made grade recommendations on an 8-step ordinal scale, supposing the exam to have been done in September. The scale, listed below, had roughly half-year intervals, its even-numbered steps representing traditional grade levels.

1. Young preschool group--below age 4
2. 4-year-old preschool group
3. Young kindergarten (pre-K)
4. Kindergarten (K)
5. Pre-first readiness class
6. First grade
7. Grade between first and second
8. Second grade

Mean scores for each measure were computed across the three main examiners, and also across four examiners including Judge F. Standard

scores by age were computed for these mean scores. At each age, boys' and girls' scores were pooled and scored on the same scales.

Results

Mean Ratings by Age

Table 2 shows mean scores on the composite rating and recommended grade for individual judges at each age, and the mean ratings across the three main judges. Table 3 shows ratings for the individual tests averaged across the main judges. Ratings on the tests and the composite were developmental ages; those for recommended grade were on the 8-step scale listed above.

Insert Tables 2 and 3 about here

The mean developmental ratings of the main judges generally fell below chronological age (CA) expectancies--that is, below a value of 4.0 at age 4, 5.0 at age 5, etc. At age 4, means lagged the age expectancy by about 2 months; by age 6 they lagged by about 7 months. Similarly, average grade recommendations were below those most common for the children's CAs. (But note that CA 6, for example, here meant children aged 5/11 to 6/1.)

Judge F knew of these lags when making her evaluations. Her mean ratings were closer to CA expectancies--only slightly for most tests at most ages, more substantially for grade recommendations.

Mean Ratings by Sex and Socioeconomic Status

Table 4 presents mean standard scores (M = 50, SD = 10) by age and by occupation group for the average of the four judges' ratings, for ages 4 to

6 years pooled. Except on the Cubes test, all the sex differences were significant, ranging in size from 5.0 to 6.0 standard score units, all favoring girls. The average boy's rating fell behind the average girl's by

---

Insert Table 4 about here

---

about 2 1/2 to 3 months. By contrast, scores for children from white-collar families--A to C on the occupation scale--did not differ significantly from those for children from blue-collar families--D to H.

Intercorrelations Among Variables

Correlations among the judges' mean scores on the five variables are presented by age in Table 5. Nearly all were moderate to high. Copy Forms and Incomplete Man ratings were highly interrelated and were the main determiners of the composite rating and grade assigned. Grade assignment

---

Insert Table 5 about here

---

and the composite rating were highly related but not automatically identical. Their lower correlation at age 4 results from a restriction in range of choices at the lower end of the grade placement scale.

Agreement Among Judges

Table 6 presents the correlations among the different judges' ratings of the same children's tests. The first row for each test gives the mean correlation across the six pairings of individual judges (at age 4, for

example, the mean of correlations between Judges A and B, A and C, A and F, B and C, B and F, and C and F).

Insert Table 6 about here

The practice of having records evaluated by more than one judge has been encouraged. The result of pooling judges' ratings is shown in the second row for each test of Table 6. For each record the scores given by two judges were averaged, these values were correlated with the average of the scores of the remaining two judges, and correlations were averaged for the three possible pairings (at age 4, A+B vs. C+F, A+C vs. B+F, and A+F vs. B+C).

Agreement between judges was highest for Cubes, the test having the simplest scaling. Correlations were intermediate and about equal for Copy Forms, Incomplete Man, and the composite score, and were lowest for recommended grade. Having two judges rate each record considerably increased agreement.

Table 7 gives another index of agreement: the mean percentage of agreement between individual judges and between pairs of judges. For the composite ratings, agreement was recorded as falling within 3 months or

---

Insert Table 7 about here

---

within 6 months. For the grade recommendation, agreement was recorded as identical or as differing by not more than one step on the scale of grade

levels listed above. (The averaged scores for pairs of judges were rounded upward to the nearest whole number.)

The judges' ratings generally showed greater agreement for the younger children than for the older ones. The composite measures showed moderate agreement between the judges. Though in theory the grade recommendations come directly--almost automatically--from the composite ratings, there was not a complete correspondence between the two; the reliability of the latter was higher than that of the former.

Retest Reliability

Judges A and F rated the records of the small sample of 4-year-olds tested for retest reliability. Each judged the first and the retest records in separate sets. The means of the individual judges' test-retest correlations were .82 and .79 respectively for composite developmental rating and grade recommended. The average developmental rating on the first testing was 50.6 months, on the second was 51.3 months, not a significant difference--in fact, slightly less of a difference than the 5 weeks that had elapsed. The average child's second developmental rating differed from his first (ignoring signs) by 1.8 months. When the judges' ratings were pooled, retest correlations between their averaged scores were .84 and .80 for developmental score and recommended grade respectively.

Relations with the Revised Gesell Preschool Examination

Table 8 shows the relations between standard scores on the screening exam and on the RGPE for all ages combined. (Findings by individual ages were very similar to those for the total sample.) Mean judges' ratings were correlated with scores on the five preschool scales and on a total

score combining the Adaptive, Verbal, Number, and Personal-Social scales. Judges' ratings showed negligible relations with the Motor scale of the

---

Insert Table 8 about here

---

full exam, low correlations with the Personal-Social, and moderate correlations with the Verbal and Number scales. As expected, judges' ratings showed highest correlations with the Adaptive scale, since both scored performance on the same tests. The judges' mean composite ratings showed about as high correlations with scores on the preschool Adaptive scale as ratings by individual judges showed with one another.

Relations with Measures at Age 8

Table 9 presents the relations of the children's composite ratings and grade recommendations with their scores on the battery of measures given 2 1/2 to 4 1/2 years later. The comparable longitudinal correlations are also shown for the RGPS Adaptive and Total (Adaptive + Verbal + Number + Personal-Social) scores. Correlations for these summary scores are presented by age group and also for ages 4 to 6 combined (via standard scores). Table 10 presents correlations for the individual tests with the outcome measures, for all ages combined. At every age a few children had incomplete records, as when teacher ratings could not be obtained. Correlations were computed using all the data available.

Insert Tables 9 and 10 about here

Though the correlation values of Table 9 show irregularities, at least in part because of small ns in some age groups, several general trends appear. First, all the screening measures except Cubes showed significant correlations with all of the follow-up measures. Even relations with 8-year-old motor skills and social maturity ranged around .35 to .40 across all ages. The highest correlations of the early test scores were with later adaptive ability, with ages-pooled values of around .55. The remaining follow-up measures--language, reading and number ability, and school competence--fell in between, with values of .40 to .45.

Second, for many of the outcome variables there appeared the expected age-to-age increase in correlation, as the time between first and second testing decreased--up to age 5 1/2. But then in most cases, at age 6 a decrease appeared in correlation with the outcome measures--often a large one. Third, as predictors of later behavior, the judges' holistic ratings and the quantitative scoring of the RGPE Adaptive scale were about equally successful. But when RGPE Verbal, Number, and Personal-Social scores were averaged with those of the Adaptive scale, predictiveness of the preschool scores was increased for several outcome measures. This effect was greatest for eventual reading, language, and number abilities.

To show the associations more concretely, for the outcome variables most relevant to the aims of the screening evaluation Table 11 gives the range of outcomes for children whose early grade placement recommendations

---

Insert Table 11 about here

---

had been relatively high or low. At each age from 4 1/2 to 6, the 25% of children having the lowest mean grade recommendations and the 25% having the highest were selected. For these initially-low and the initially-high groups, quartile distributions were determined for each outcome measure. For example, of all the 4 1/2- to 6-year-olds whose recommendation of grade placement fell in the lowest 25% of scores for their age, 30 were tested on the Adaptive cluster at age 8. Of these 30, 53% fell in the lowest quartile of 8-year-old scores, Q_1, while 7% fell in the highest quartile, Q_4. Differences in outcomes were marked between the groups having high and low placement recommendations.

Discussion

Mean scores

Mean developmental ratings by the main judges were lower than corresponding CAs by from two to seven months, and mean grade placement recommendations were lower than those customary for the ages. Several factors may have played a role in this. (a) Age groupings sound misleading. The "6-year-olds," for example, were aged 5/11 to 6/1. More findings of immature-for-first-grade would be expected than in a sample of 6-year-olds who ranged from 6/0 to 6/11. (In a locale with a July-birthday cut-off for school entrance--such as the state of Missouri--100% of these 6-year-olds would ordinarily be ineligible for first grade.) (b) In actual placement situations, where practical considerations may limit options for placement, experienced developmental examiners regularly report finding about one third of the children in most unselected groups should be considered for extra time at an early grade level. In this

research task, hypothetical and open ended, the judges may perhaps have been extra vigilant--"super-Gesellian"--against overplacement. (They reported wanting much more information about both child and school before making real-life placement recommendations.)

(c) Probably most important, the judges appeared to work from a conceptual model different from the usual psychometric assumption. For a global variable of developmental maturity, their conception results in a distribution of children's developmental ages more like a J-curve than a bell-shaped one. Some children score much younger than their life ages, then more and more children are found as scores move up to where developmental age and life age correspond--lots of 5-year-olds should score as developmentally 5. But a kind of wall appears just beyond that point, so that very few 5-year-olds score developmentally as high as 5 1/2. Above-CA scores may readily occur for particular items, fewer occur for individual tests, fewer still for overall ratings: the level at which the part functions get put together and organized rarely exceeds the child's life age. In this model it is unlikely that the mean developmental score at an age will match the CA, though the modal score may do so. In actuality, the mode also often falls below CA. Ratings can evidently be somewhat modified to fit the demands of a particular situation. The records were rescored by Judge F, who knew that the other judges' scores had fallen below CA expectancies and was encouraged toward liberality in rating. The lag found in this judge's mean developmental scores was about half that of the other judges, and her grade recommendations were closer to traditional placements. What seems striking is not that her global ratings

veered toward somewhat higher values but that they did not go further. In a more psychometric approach, scales might be recalibrated fairly easily; but an experienced rater trying to assess individual children could not, in good conscience, score them higher than she did.

Assigning developmental ages to tests involves one set of variables, within the child; recommending a grade placement involves two: the child's characteristics and the (real or hypothetical) school's structure--curriculum, classrooms and spaces available, teachers staffing the classes. It seems unsurprising that Judge F's mean placement recommendations could differ more from those of the other judges than her developmental-age scores, if what constitutes readiness for kindergarten, say, depends in part on how one pictures the standard kindergarten.

Sex and Socioeconomic Differences

Except on the Cubes test, judges' mean ratings for girls consistently exceeded those for boys, by the equivalent of something less than 3 months of growth. Significant differences did not appear between occupational groupings. These findings differ from the typical observation with older children and adolescents that boys tend to perform as well as or better than girls in spatial-perceptual-motoric tasks, and that differences by occupational grouping are very common. The judges knew the children's sexes but did not know their SES. That the sex difference was not a result of preconceptions on the judges' part was indicated by parallel findings of about the same size when the children's records were scored rather than rated (Walker, in press): the preschool girls were the more mature in virtually all the areas tested. The RGPE also showed no SES difference for

four of its five scales: Motor, Adaptive, Number, and Personal-Social. A small but significant SES difference did appear on the Verbal scale.

In an ethnically diverse sample of 48 kindergarten children, Lichtenstein (1989) reported parallel findings for the Gesell screening exam: the girls scored slightly higher than the boys but there were no subgroup differences by SES (nor by language status nor ethnic background), though on tests from the K-ABC for the same children, the boys scored higher than the girls and significant subgroup differences were found by SES, language status, and ethnic background.

Reliability

Considered as global ratings of complex variables, the correlation among the judges' scores was moderate to high: roughly .75 to .85 for the composite ratings, .65 to .75 for the grade recommendations. Considered as a basis for making placement decisions, their level of agreement arouses concern. Though concensus was substantial, disagreement among experienced examiners was not infrequent. In his study of Gesell screening exam scores in kindergarten children, Lichtenstein (1989) reported an inter-rater correlation of .77 for grade placement recommendations, as well as test-retest reliability of .74 over one to five weeks. Both his values were close to findings here.

Other preschool tests and screening instruments, made up of simple, clearly defined items, report interrater reliability values much higher than those in this study. The Denver Developmental Screening Test (DDST; Frankenburg, Camp, Van Natta, Demersseman & Voorhees, 1971), the Early Screening Inventory (ESI; Meisels, Wiske, & Tivnan, 1984), and the

Minneapolis Preschool Screening Instrument (MPSI; Lichtenstein & Ireton, 1984) report median interobserver values from .91 to .97. Despite this, typically reported test-retest and internal consistency measures of reliability are comparable to the Gesell Screening test-retest correlation values. For example, for the individual tests of the WPPSI, Wechsler (1967) reported mean internal consistency correlations of .77 to .87 across ages 4 to 6 1/2 for his norm sample, and a median test-retest correlation of .69 for a sample of 50 5-year-olds. The McCarthy Scales of Children's Abilities (McCarthy, 1972) reports median test-retest reliability of .57. The DDST (Frankenburg et al, 1971) reports test-retest values around .78, the ESI (Mesiels et al, 1984) a test-retest value of .82. That is, the Gesell screening interjudge correlations are of the same order as the test-retest and internal consistency measures of reliability reported in other studies, and their comparatively low values do not appear to result in lower test-retest values.

The judges in our study reported experiencing three handicaps in scoring. (a) They were dealing with records alone, not observing children, and could evaluate only products, not children's processes of working. (b) The exams did not include the verbal tests and occasionally lacked the Cubes test. (c) Recommendations were for placement in hypothetical classrooms. The lower correlations for grade recommendations presumably arose from differences in picturing of classes to match different maturities. If the judges had themselves tested the children, for placement in a familiar school, their agreement might have been greater.

The practice has been encouraged of having records evaluated by more than one judge--for example, having a second judge review recommendations or obtaining independent ratings of the same records, with a conference or a third judge to resolve disagreements. Pooling ratings of two judges for each record considerably increased reliability of the scores.

Individual Measures

Among the individual tests rated, Copy Forms and Incomplete Man were highly intercorrelated and showed similar relations to most other variables; the Cubes test stood apart from the other two. Judges showed relatively good agreement in rating Cubes, because ratings were tied to a simple scale. But judges gave relatively little weight to Cubes in making composite ratings or grade recommendations; its curve of mean scores by age was the shallowest among the three tests; it showed the lowest correlations with the RGPE; and it had little predictive value for any outcome measures. The judges reported giving lesser weighting to Cubes because some test records were not well recorded.

Outcomes

When decisions are made about a group of children--these need extra time at an early grade level, these do not--it is reasonable later to ask, were the decisions right? What was the hit rate? If we had a sound measure of what constituted a good outcome for each child, we would be in a good position to answer the question. Instead, researchers must impose simplified classifications on outcomes in which children show complex variation. For example, high achievement test scores are desirable outcomes but used alone achievement is an inadequate criterion of

successful placement, if success means determining a setting in which the child is both comfortable and productive. Some high-achieving children are out of their depth in a grade that matches their achievement alone.

The validity of the screening tests, as shown by their power to predict later characteristics, appeared in correlations that rose from age 4 to age 5 1/2 to reach about .70 with the tests having the highest content similarity (the Adaptive follow-up tests). The screening ratings also reached peak correlations of about .60 with plausibly related skills (Fine-motor) and about .45 to .50 with school-related abilities (Reading, Number, School Competence). Over a three-year age span, these are substantial relations. If it were not for the repeated drop in correlations at age 6, the developmental evidence would be clearer.

To date we have discovered no convincing reason for this decrease at 6, which appeared also in the correlations with the RGPE. The obvious explanation of a restricted range of variation--a ceiling effect--appears not to hold. (The SD does not drop at 6, scatter diagrams show no flattening, and there was no theoretical ceiling to the ratings by the judges, who were experienced in rating records of 7- and 8-year-olds.) The only confounding variable effect found has been with SES: the correlation between 6-year-old recommended grade and 8-year-old Adaptive score was .72 --about as expected--for the 23 children from blue-collar homes, but only .21 for the 24 from white-collar homes. But no such SES effect appeared at age 5 1/2, and the effect is presumably a chance sampling variation.

The comparison of extreme groups gives a realistic picture of the outcomes of the testing without resorting to pass/fail simplifications.

Children who initially clearly scored high on the screening exam and those who scored low differed sharply from one another on a variety of measures some 2 1/2 to 4 years later. The differences in outcomes for the contrasted groups argue against the often-heard position that by third grade any effects of earlier differences in maturity have disappeared.

The conclusions one draws depend on what side of the fence one is standing on. The findings show the screening measures to have moderate reliability and predictive power. They also show that experienced judges sometimes differ in assessing a child's developmental level and recommending grade placement. And on any of the outcome measures at least a few children whose early scores were immature did well three or four years later and a few whose early scores were mature later scored poorly.

The research reported here does not touch directly on the use of transitional grades or the consequences of a slower or faster course in school—the purpose of giving a screening evaluation in the first place. It does have implications for practice for those who use the Gesell screening evaluations to make placement recommendations. (a) It would be well for such recommendations to be based on a record evaluated by more than one rater. (b) The recommendation should consider other sources of information, such as the views of previous teachers and of parents. (c) Recognizing that disagreement among experienced examiners does occur and that growth sometimes brings changes in relative status, an examiner would do better to be open rather than dogmatic in stance. Evaluation ideally should be an ongoing process, with choice of the best setting for a child open to change as the child grows.

References

Ames, L. B., Gillespie, C., Haines, J., & Ilg, F. L. (1979). The child from one to six. New York: Harper & Row.

Doll, E. A. (1965). Vineland Social Maturity Scale. Circle Pines, MN: American Guidance Service.

Frankenburg, W. K., Camp, B. W., Van Natta, P. A., Demersseman, J. A., & Voorhees, S. F. (1971). Reliability and stability of the Denver Developmental Screening test. Child Development, 42, 1315-1325.

Gesell, A. (1925). The mental growth of the preschool child. New York: Macmillan.

Gesell, A. & Amatruda, C. S. (1941). Developmental diagnosis. New York: Hoeber.

Gesell, A., Halverson, H. M., Thompson, H., Ilg, F. L., Castner, B. M., Ames, L. B., & Amatruda, C. S. (1940). The first five years of life. New York: Harper & Row.

Gesell, A. & Thompson, H. (1938) The psychology of early growth. New York: Macmillan.

Haines, J., Ames, L. B., & Gillespie, C. (1980). The Gesell Preschool Test manual. Lumberville, PA: Modern Learning Press.

Ilg, F. L. (1985). Scoring notes for the Developmental Examination. New Haven: Gesell Institute of Human Development.

Ilg, F. L., & Ames, L. B. (1964). School readiness. New York: Harper & Row.

Jastak, J. F., Bijou, S. W., & Jastak, S. R. (1965). Wide Range Achievement Test. Wilmington, DE: Guidance Associates of Delaware.

Lichtenstein, R. (1989). Psychometric characteristics and appropriate use of the Gesell School Readiness Screening Test. Manuscript submitted for publication.

Lichtenstein, R., & Ireton, H. (1984). Preschool screening: Identifying young children with developmental and educational problems. Orlando, FL: Grune & Stratton.

McCarthy, D. (1972). McCarthy Scales of Children's Abilities. New York: Psychological Corporation.

Meisels, S. J., Wiske, M. S., & Tivnan, T. (1984). Predicting school performance with the Early Screening Inventory. Psychology in the Schools, 21, 25-33.

Walker, R. N. (1989). The Revised Gesell Preschool Examination and later competence. Manuscript submitted for publication.

Walker, R. N. (in press). The Revised Gesell Developmental Examination: Manual. New York: Programs for Education.

Wechsler, D. (1967). Manual for the Wechsler Preschool and Primary Scale of Intelligence. New York: Psychological Corporation.

Table 1

Percentage Distribution of Fathers' Occupations[a] for Normative Sample and for Retested Subsample, Compared with 1970 U.S. Census Data

Occupation	1970 Census[b]	Norm sample	Retested subsample
A. Professional and technical	14.9	15.0	15.4
B. Managers, officials, proprietors	15.2	15.0	14.8
C. Clerical and sales	12.5	12.5	12.6
D. Craftsmen and foremen	21.1	21.3	29.7
E. Operatives	19.4	18.8	13.2
F. Farmers, farm foremen, farm laborers	5.1	5.0	4.4
G. Service workers	6.0	6.3	7.1
H. Nonfarm laborers	6.1	6.3	2.7
A-C. Total "white collar"	42.5	42.5	42.9
D-H. Total "blue collar"	57.5	57.5	57.1

[a]Mother's occupation used where father's occupation not given.

[b]Distribution of employed males, 20 years and over, 1970 U.S. Census, with the "not reported" category (4.6%) spread proportionally across the remaining categories.

Table 2

Means and Standard Deviations by Age for Individual Judges' Composite Ratings and Grade Recommendations

Judge		Composite rating					Recommended grade[a]				
		Age					Age				
		4	4½	5	5½	6	4	4½	5	5½	6
A	Mean	3.7	4.2	4.6	--	--	1.4	2.1	3.0	--	--
	SD	0.4	0.3	0.4	--	--	0.5	0.6	0.8	--	--
B	Mean	3.9	4.3	--	5.1	--	1.6	2.2	--	3.7	--
	SD	0.5	0.4	--	0.4	--	0.6	0.6	--	0.6	--
C	Mean	3.7	4.2	--	--	5.4	1.4	2.3	--	--	4.7
	SD	0.4	0.4	--	--	0.4	0.5	0.8	--	--	0.9
D	Mean	--	--	4.4	4.8	5.2	--	--	2.7	3.3	4.3
	SD	--	--	0.4	0.4	0.4	--	--	0.8	0.8	0.8
E	Mean	--	--	4.8	5.1	5.5	--	--	3.3	3.9	4.6
	SD	--	--	0.3	0.4	0.4	--	--	0.7	0.8	0.7
F	Mean	4.0	4.4	4.8	5.1	5.7	1.7	2.2	3.3	4.2	5.1
	SD	0.4	0.3	0.3	0.3	0.4	0.5	0.7	0.9	0.8	0.8
Mean,	Mean	3.8	4.2	4.6	5.0	5.4	1.5	2.2	3.0	3.6	4.5
A to E	SD	0.4	0.4	0.3	0.3	0.4	0.5	0.6	0.6	0.6	0.8

Note. n = 80 in each cell. [a]See [illegible] for scale of grade levels.

Table 3

Means and Standard Deviations of Ratings of Individual Tests by Three Main Judges, Averaged, and by Judge F

Score			Age				
			4	4½	5	5½	6
Copy Forms							
	Main judges	Mean	3.8	4.2	4.6	5.0	5.4
		SD	0.4	0.4	0.4	0.4	0.4
	Judge F	Mean	4.0	4.4	4.8	5.1	5.7
		SD	0.4	0.4	0.4	0.3	0.4
Incomplete Man							
	Main judges	Mean	3.7	4.2	4.6	5.0	5.3
		SD	0.4	0.4	0.4	0.4	0.4
	Judge F	Mean	4.0	4.5	4.9	5.1	5.6
		SD	0.4	0.3	0.3	0.3	0.4
Cubes							
	Main judges	Mean	4.1	4.5	4.8	5.1	5.3
		SD	0.6	0.6	0.6	0.6	0.6
	Judge F	Mean	4.3	4.6	4.9	5.1	5.3
		SD	0.5	0.6	0.6	0.6	0.6

Note. ns involving Cubes range from 75 to 79 at individual ages; all other ns = 80 at each age.

Table 4

Mean Standard Scores by Sex and by Paternal Occupation for Judges' Ratings

Category	n	Standard score				
		Copy forms	Incomplete man	Cubes[a]	Composite score	Grade assigned
By Sex						
Boys	200	47.5	46.9	49.2	47.2	47.1
Girls	200	52.5*	53.1*	50.8	52.8*	52.9*
By Paternal Occupation						
A. Professional, technical	60	52.5	51.4	50.7	52.2	51.6
B. Managers, officials	60	48.8	48.7	50.3	48.8	49.3
C. Clerical, sales	50	50.2	50.7	51.8	50.5	50.8
D. Craftsmen, foremen	85	51.3	51.0	47.9	50.7	50.7
E. Operatives	75	48.5	49.4	48.8	48.7	49.2
F. Farmers	20	47.7	48.1	49.8	48.7	47.9
G. Service workers	25	50.6	49.8	52.7	50.3	49.4
H. Nonfarm laborers	25	47.6	48.4	52.2	48.5	48.0
A-C. Total white collar	170	50.5	50.2	50.9	50.5	50.5
D-E. Total blue collar	230	49.6	49.8	49.4	49.6	49.4

[a] n for Cubes test slightly lower than listed in some rows: 15 of the 400 records were inadequately recorded for blind rating. *$p < .01$.

Table 5

Correlations Among Mean Judges' Ratings

Age group	Inc. Man vs. Copy Forms	Cubes vs. Copy Forms	Cubes vs. Inc. Man	Composite rating vs. Copy Forms	Composite rating vs. Inc. Man	Composite rating vs. Cubes	Grade assigned vs. Copy Forms	Grade assigned vs. Inc. Man	Grade assigned vs. Cubes	Grade assigned vs. Composite
4	.85	.52	.53	.93	.96	.64	.83	.89	.67	.92
4½	.88	.41	.37	.95	.96	.51	.92	.91	.48	.95
5	.90	.35	.34	.94	.96	.49	.94	.95	.50	.98
5½	.82	.32	.37	.91	.93	.54	.90	.92	.50	.97
6	.85	.49	.45	.94	.94	.63	.93	.94	.60	.98
All	.85	.40	.39	.92	.94	.55	.88	.91	.53	.95

Note. Inc. Man = Incomplete Man test. ns involving Cubes range from 75 to 79; all other ns = 80. Correlations for ages pooled were computed through standard scores. $p < .05$ for all correlations.

Table 6

Agreement Among Judges' Ratings: Mean Correlations Across All Pairings of Individual Judges and Across Pairings of Paired Judges

Measure	Age				
	4	4½	5	5½	6
Copy Forms					
Individual judges	.82	.82	.78	.73	.73
Pairs of judges	.91	.91	.88	.84	.84
Incomplete Man					
Individual judges	.82	.80	.76	.80	.72
Pairs of judges	.90	.89	.86	.88	.83
Cubes					
Individual judges	.88	.87	.83	.92	.87
Pairs of judges	.94	.93	.91	.96	.93
Composite rating					
Individual judges	.85	.82	.74	.79	.78
Pairs of judges	.92	.90	.85	.88	.87
Grade recommended					
Individual judges	.64	.67	.70	.68	.74
Pairs of judges	.78	.80	.82	.81	.85

Note. ns involving Cubes range from 75 to 79 at individual ages; all other ns = 80 at each age.

Table 7

Agreement Among Judges' Ratings: Mean Percentage of Agreement for All Pairings of Individual Judges and for Average Scores of Pairs of Judges

Measure	Age				
	4	4½	5	5½	6
Composite ratings					
Agree within 3 months					
Individual judges	74	86	70	78	71
Pairs of judges	80	90	75	81	78
Agree wit 6 months					
Indivi al judges	94	97	92	93	92
Pairs of judges	99	100	98	98	98
Grade recommended					
Agree identically					
Individual judges	76	71	56	60	56
Pairs of judges	86	81	77	71	70
Agree within 1 level					
Individual judges	100	100	96	96	90
Pairs of judges	100	100	98	98	98

Note. n = 80 at each age.

Table 8

Correlations Between Mean Judges' Ratings and Scores on Scales of the Revised Gesell Preschool Examination, Ages 4 to 6 Years Combined

Preschool Scale	Judges' rating				
	Copy Forms	Incomplete Man	Cubes	Composite rating	Grade recommended
Motor	.15*	.20*	.07	.18*	.16*
Adaptive	.79*	.76*	.53*	.82*	.79*
Verbal	.36*	.38*	.16*	.37*	.35*
Number	.40*	.38*	.22*	.38*	.37*
Personal-social	.28*	.27*	.12*	.28*	.27*
Total score[a]	.63*	.61*	.36*	.64*	.62*

Note. n = 385 for Cubes; n = 400 for all other cells. Correlations computed through standard scores by age. [a]Total score = sum of standard scores for individual scales omitting Motor. *$p < .01$.

Table 9

Correlations of Outcome Scores with Judges' Composite and Grade Ratings and with Adaptive and Total Scores on the Revised Gesell Preschool Examination

8-year-old measure	Age	Judges' ratings		Preschool Exam	
		Composite rating	Grade recommended	Adaptive score	Total score[a]
Gross motor	4	.33	.35	.37	.54*
	4½	.36	.34	.26	.16
	5	.39	.38*	.33*	.46**
	5½	.39*	.36*	.41**	.45**
	6	.50**	.53**	.30*	.28
	All	.42**	.42**	.33**	.36**
Fine motor	4	.45	.42	.18	.13
	4½	.21	.23	.30	.33
	5	.38*	.37*	.27	.38*
	5½	.62**	.66**	.56**	.60**
	6	.20	.23	.16	.21
	All	.38**	.39**	.30**	.35**
Adaptive	4	.27	.18	.37	.59**
	4½	.47**	.51**	.52**	.56**
	5	.60**	.62**	.55**	.61**
	5½	.73**	.75**	.68**	.67**
	6	.50**	.50**	.46**	.47**
	All	.56**	.57**	.54**	.58**

(table continues)

Table 9, continued

8-year-old measure	Age	Judges' ratings		Preschool Exam	
		Composite rating	Grade recommended	Adaptive score	Total score[a]
Language	4	.36	.31	.31	.39
	4½	.46**	.47**	.50**	.44*
	5	.51**	.54**	.42**	.63**
	5½	.44**	.35*	.62**	.71**
	6	.41**	.41**	.44**	.53**
	All	.44**	.43**	.47**	.56**
Reading	4	.33	.30	.32	.38
	4½	.44*	.46*	.50**	.42*
	5	.49**	.51**	.40*	.61**
	5½	.51**	.41**	.71**	.74**
	6	.35*	.37*	.36*	.42*
	All	.43**	.42**	.45**	.52**
Number	4	.29	.21	.28	.45
	4½	.52**	.47**	.51**	.51**
	5	.45**	.48**	.32*	.48**
	5½	.52**	.43**	.64**	.60**
	6	.55**	.59**	.49**	.51**
	All	.48**	.46**	.46**	.52**

(table continues)

Table 9, continued

8-year-old measure	Age	Judges' ratings		Preschool Exam	
		Composite rating	Grade recommended	Adaptive score	Total score[a]
Social maturity	4	.41	.35	.38	.58**
	4½	.21	.25	.38*	.36*
	5	.32*	.30*	.26	.34
	5½	.53**	.55**	.55**	.65**
	6	.38*	.43**	.33*	.44**
	All	.36**	.37**	.35**	.44**
Competence	4	.35	.36	.25	.38
	4½	.45*	.43*	.46*	.45*
	5	.45**	.46**	.31*	.46**
	5½	.49**	.46**	.44**	.55**
	6	.40**	.44**	.51**	.56**
	All	.43**	.44**	.41**	.49**

Note. n was 19-20 at age 4, 27-30 at 4½, 37-40 at 5, 39-40 at 5½, 47-48 at 6, 166-182 for the total sample. Correlations for ages combined computed through standard scores by age. [a]Total Preschool Exam score = sum of Adaptive + Verbal + Number + Personal-social standard scores.

*$p < .05$. **$p < .01$.

Table 10

Correlations of Individual Preschool Measures with Outcome Scores for Ages 4 to 6 Years, Pooled

8-year-old measure	n	Judges' preschool ratings		
		Copy Forms	Incomplete Man	Cubes
Group measures				
Gross motor	176	.35**	.42**	.22**
Fine motor	176	.35**	.39**	.17*
Adaptive	175	.56**	.49**	.19*
Verbal	176	.44**	.42**	.23**
Reading	176	.43**	.39**	.24**
Number	176	.51**	.43**	.27**
Social maturity	182	.35**	.37**	.13
Competence	172	.45**	.45**	.13

Note. ns for correlations involving Cubes are 6 less than listed ns.
*$p < .05$. **$p < .01$.

Table 11

Percentage Distributions on Four Outcome Measures for Children in the Lowest 25% or the Highest 25% of Judges' Grade Placement Recommendations at Ages 4½ to 6 Years

8-year-old measure	Judges' Rating at 4½-6	n	8-year-old status			
			Q_1	Q_2	Q_3	Q_4
Adaptive	Low	30	53	27	13	7
	High	40	3	18	30	50
Reading	Low	30	47	27	17	10
	High	40	5	28	28	40
Number	Low	30	47	37	13	3
	High	40	3	20	40	38
Competence	Low	30	47	30	13	10
	High	40	8	23	30	40